Virus Evolution, 2024, **10(1)**, veae071
DOI: https://doi.org/10.1093/ve/veae071
Advance Access Publication 2 September 2024
Research Article

SARS-CoV-2 variant replacement constrains vaccine-specific viral diversification

Bethany L. Dearlove[1,2], Anthony C. Fries[3], Nusrat J. Epsi[2,4], Stephanie A. Richard[2,4], Anuradha Ganesan[2,4,5], Nikhil Huprikar[5], David A. Lindholm[6,7], Katrin Mende[2,4,7], Rhonda E. Colombo[2,4,6,8], Christopher Colombo[8], Hongjun Bai[1,2], Derek T. Larson[9], Evan C. Ewers[9], Tahaniyat Lalani[2,4,10], Alfred G. Smith[10], Catherine M. Berjohn[4,6,11], Ryan C. Maves[12], Milissa U. Jones[13], David Saunders[6], Carlos J. Maldonado[14], Rupal M. Mody[15], Samantha E. Bazan[16], David R. Tribble[4], Timothy Burgess[4], Mark P. Simons[4], Brian K. Agan[2,4], Simon D. Pollett[2,4], Morgane Rolland[1,2]

[1]US Military HIV Research Program, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910, United States
[2]Henry M Jackson Foundation for the Advancement of Military Medicine, Inc., 6720A Rockledge Drive, Bethesda, MD 20817, United States
[3]The Applied Technology and Genomics (PHT) Division, US Air Force School of Aerospace Medicine, 2510 5th St, Dayton, OH 45433, United States
[4]Infectious Disease Clinical Research Program, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, United States
[5]Division of Infectious Diseases, Walter Reed National Military Medical Center, 8901 Rockville Pike, Bethesda, MD 20889, United States
[6]Department of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, United States
[7]Division of Infectious Diseases, Brooke Army Medical Center, 3551 Roger Brooke Drive, San Antonio, TX 78234, United States
[8]Division of Infectious Diseases, Madigan Army Medical Center, 9040 Jackson Avenue, Tacoma, WA 98431, United States
[9]Division of Infectious Diseases, Alexander T. Augusta Military Medical Center, 9300 DeWitt Loop, Fort Belvoir, VA 22060, United States
[10]Division of Infectious Diseases, Naval Medical Center Portsmouth, 620 John Paul Jones Circle, Portsmouth, VA 23708, United States
[11]Infectious Diseases and Internal Medicine, Naval Medical Center San Diego, 34800 Bob Wilson Drive, San Diego, CA 92134, United States
[12]Sections of Infectious Diseases and Critical Care Medicine, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, United States
[13]Department of Pediatrics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, United States
[14]Department of Clinical Investigation, Womack Army Medical Center, 2817 Rock Merritt Avenue, Fort Liberty, NC, United States
[15]Division of Infectious Diseases, William Beaumont Army Medical Center, 18511 Highlander Medics Street, El Paso, TX 79918, United States
[16]Department of Primary Care, Carl R. Darnall Army Medical Center, 590 Medical Center Road, Fort Cavazos, TX 76544, United States

Corresponding author. Morgane Rolland, mrolland@hivresearch.org

Abstract

Coronavirus disease 2019 (COVID-19) vaccine breakthrough infections have been important for all circulating severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variant periods, but the contribution of vaccine-specific SARS-CoV-2 viral diversification to vaccine failure remains unclear. This study analyzed 595 SARS-CoV-2 sequences collected from the Military Health System beneficiaries between December 2020 and April 2022 to investigate the impact of vaccination on viral diversity. By comparing sequences based on the vaccination status of the participant, we found limited evidence indicating that vaccination was associated with increased viral diversity in the SARS-CoV-2 spike, and we show little to no evidence of a substantial sieve effect within major variants; rather, we show that rapid variant replacement constrained intragenotype COVID-19 vaccine strain immune escape. These data suggest that, during past and perhaps future periods of rapid SARS-CoV-2 variant replacement, vaccine-mediated effects were subsumed with other drivers of viral diversity due to the massive scale of infections and vaccinations that occurred in a short time frame. However, our results also highlight some limitations of using sieve analysis methods outside of placebo-controlled clinical trials.

Keywords: SARS-CoV-2; vaccine breakthrough; sieve analysis; variants.

Introduction

There have been multiple waves of coronavirus disease 2019 (COVID-19) cases, and since December 2020, the distribution of effective vaccines has been key to limiting the impact of COVID-19. While the vaccines have been highly effective in preventing symptomatic disease and severe illness, breakthrough infections can still occur in vaccinated individuals (Haas et al. 2021, Hacisuleyman et al. 2021, Thompson et al. 2021). Over time, a succession of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants has spread, with each variant being more distant from the original vaccine spike gene inserts, alongside waning vaccine efficacy (VE) against symptomatic infection (Mascola et al. 2021,

© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site–for further information please contact journals.permissions@oup.com.

Tartof et al. 2021, Cao et al. 2022, Feikin et al. 2022). Cao *et al*. previously showed that VE declines with genetic distance from the receptor domain of the vaccine strain (Cao et al. 2022).

One way to understand the genetic consequences of vaccination on the virus is sieve analysis (Rolland and Gilbert 2021). This compares sequences sampled from vaccine breakthrough infections to those sampled from placebo recipients in a VE trial. One key hypothesis is that vaccines are expected to preferentially block viruses most closely related to the vaccine insert, which for most vaccines was based on the spike gene of the earliest available sequence, Wuhan-Hu-1 [Global Initiative on Sharing All Influenza Data (GISAID) accession: EPI_ISL_402125, GenBank accession: NC_045512.2], and thus sequences from vaccinated individuals will be more divergent from the vaccine sequence. This could be through the accumulation of more substitutions across a whole genome or gene or by increased diversity at specific sites. In particular, under this hypothesis, it would be expected that vaccine-mismatched residues would fall at contact sites for spike-specific antibodies, allowing for immune escape (Rolland and Gilbert 2021). Sieve analyses are normally conducted within clinical trial settings to investigate signals of sieve effects in vaccine breakthrough viruses relative to a placebo group as a control (Gilbert et al. 2008, Rolland et al. 2011, Edlefsen et al. 2015). Magaret and colleagues (including some of us) previously showed a sieve effect in the ENSEMBLE randomized, placebo-controlled Phase 3 trial (NCT04505722), which had shown a VE of 56% for the single-dose Ad26.COV2.S against moderate to severe–critical COVID-19 (Magaret et al. 2024). The analysis of SARS-CoV-2 spike sequences from 484 vaccine and 1067 placebo recipients, who were diagnosed with COVID-19, showed that VE was reduced against the Lambda variant in Latin America.

Although there have been efforts to associate vaccination status with surveillance sequence data in public repositories, the data remain sparse and difficult to interpret, especially considering the varying vaccination platforms and boosting regimes (Marques et al. 2021). In this study, we took advantage of the known vaccination history, including number of doses, type, and date administered, in the Epidemiology, Immunology, and Clinical Characteristics of Emerging Infectious Diseases with Pandemic Potential (EPICC) study. We investigated whether there was evidence of vaccine-mediated pressure on the SARS-CoV-2 sequences.

Materials and methods

Study population and setting

US Military Health System beneficiaries with a history of SARS-CoV-2 infection who were tested for SARS-CoV-2 and/or who received COVID-19 vaccination were eligible for enrollment into the EPICC study, a prospective, longitudinal observational cohort study. The EPICC cohort has been described elsewhere (Richard et al. 2021, 2023, Epsi et al. 2023a). Briefly, participants completed surveys over a 1-year period, with further clinical information abstracted from their electronic medical records. Repeated biospecimens were collected over 1 year, including blood and upper respiratory tract swabs (Richard et al. 2021) (see Supplementary Table S1 for a summary of study procedures).

Enrollment occurred at 10 military treatment facilities (MTFs): Brooke Army Medical Center (BAMC), Alexander T. Augusta Military Medical Center (ATAMMC), Naval Medical Center Portsmouth (NMCP), Walter Reed National Military Medical Center (WRNMC), Carl R. Darnall Army Medical Center (CRDAMC), William Beaumont Army Medical Center (WBAMC), and Womack Army Medical Center (WAMC) are in the South census region, and Madigan Army Medical Center (MAMC), Naval Medical Center San Diego (NMCSD), and Tripler Army Medical Center (TAMC) are in the West census region. This analysis leveraged those EPICC participants with a documented history of SARS-CoV-2 infection, with available viral sequence data, and with infections occurring in the era of COVID-19 vaccine availability in the Military Health System (see Supplementary Fig. S1).

Detection of SARS-CoV-2 infection

SARS-CoV-2 infections were detected using a positive clinical laboratory polymerase chain reaction (PCR) test performed at the enrolling clinical site or on an upper respiratory swab collected as part of the EPICC study, as described in Richard et al. (2021). Different PCR assays were used at the MTFs, and the SARS-CoV-2 (2019-nCoV) Centers for Disease Control and Prevention (CDC) qPCR Probe Assay research-use-only kit (Integrated DNA Technologies, Coralville, IA) was used for testing specimens collected as part of the EPICC study procedures. This qPCR assay used two SARS-CoV-2 nucleocapsid (N) gene (N1 and N2) targets and a human RNase P gene (RP) control. SARS-CoV-2 positivity was defined using a cycle threshold value of <40 for both N1 and N2 gene targets.

Ascertainment of vaccination status and other independent variables

Demographic information, baseline health, COVID-19 vaccination history, and other characteristics were obtained from surveys and abstraction from the electronic medical record (case report forms and abstraction from the Military Health System Data Repository) as previously described (Epsi et al. 2023b). For the purposes of this study, we defined "fully vaccinated" before infection as having received two mRNA COVID-19 vaccine doses or one viral-vectored vaccine dose >2 weeks before their first positive SARS-CoV-2 test date.

Viral sequencing methods

Whole viral genome sequencing on SARS-CoV-2-positive swabs and residual clinically collected swabs was performed as described in prior EPICC papers (Lusvarghi et al. 2022, Richard et al. 2022, Wang et al. 2022). Briefly, extracted SARS-CoV-2 RNA from PCR-positive specimens was sequenced using a 1200-bp amplicon tiling strategy (Freed et al. 2020). NexteraXT library kits (Illumina Inc., San Diego, CA) were used to prepare amplified products for sequencing. Libraries were run on the Illumina NextSeq 550 sequencing platform. BBMap v8.86 and iVar v1.2.2 tools were used for genome assembly. Sequences were classified into Pango lineages using pangolin v4.3 with data version 1.22 (O'Toole et al. 2021).

EPICC sequence selection

In the few cases where a participant had more than one sample sequenced with high-quality genome coverage, we downsampled to one sequence. In all cases, the longitudinal sequences were sampled within 6 days of the first available sequence, and since some follow-up sequencing was due to low quality, we retained the sequence with the highest quality represented by the highest proportion of bases with coverage in the spike gene and highest percentage of sites with >20× coverage. The GenBank accession numbers of the final sequence dataset are OR611156–OR611708 and PP378952–PP379010.

Context sequences

Context information about the circulation of variants in the South and West census regions of the USA was downloaded from GISAID (Khare et al. 2021) using the metadata summary package available as of 28 April 2022 and by filtering to retain sequences after 1 December 2020 to align with vaccination availability in the USA. Data were deduplicated by name, keeping the earliest accession, and sequences with >5% gaps and missing full dates were removed. Sequences assigned to a variant lineage that had a date prior to that variant's emergence were also removed, under the assumption that these sequences were much more diverged than expected and rather reflected metadata errors. Thresholds followed those used by NextStrain (available at https://github.com/nextstrain/ncov/blob/master/defaults/clade_emergence_dates.tsv), except for Iota, which was adjusted earlier: Alpha, 20 September 2020; Beta, 10 August 2020; Gamma, 29 October 2020; Delta, 30 October 2020; Kappa, 30 October 2020; Epsilon, 3 August 2020; Eta, 21 November 2020; Iota, 1 November 2020; Mu, 5 January 2021; Lambda, 5 January 2021; Theta, 10 January 2021; and Omicron, 1 September 2021. Accessions included in the analysis are available on GISAID via gisaid.org/EPI_SET_230903xp.

Emergence of Alpha, Delta, and Omicron across groups

We fit independent logistic models to estimate the proportion of cases attributable to Alpha, Delta, and Omicron over time, including vaccination group as a covariate. The models take the form:

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right]=\alpha+\beta_1X_1+\beta_2X_2,$$

where $Y=1$ for samples matching the variant of concern (VOC) being compared and $Y=0$ for any other lineage, X_1 is the date of collection, and $X_2=0$ for unvaccinated and $X_2=1$ for vaccine breakthrough. Thus, for the Alpha comparison, a sequence has $Y=1$ if it is classified as Alpha, but $Y=0$, if it is any other lineage. Consequently, this analysis does not consider the underlying background diversity from which a variant emerged. Since logistic regression can only estimate the emergence until peak, data were restricted to before the date on which the World Health Organization (WHO) announced the next major VOC: before 11 May 2021 (date of Delta emergence announcement) for the Alpha analysis and before 26 November 2021 (date of Omicron emergence announcement) for the Delta analysis.

Sequence and phylogenetic analyses

Sequences were aligned to Wuhan-Hu-1 (GISAID Accession: EPI_ISL_402125, GenBank accession: NC_045512.2) with Mafft v7.487 using the add_fragments option (Katoh and Standley 2016). Genes were extracted using the coordinates relative to Wuhan-Hu-1 and translated.

Phylogenetic trees were reconstructed from nucleotide genome and spike sequences using IQ-TREE v2.1.3 under the best-fitting substitution model from ModelFinder (Nguyen et al. 2014, Kalyaanamoorthy et al. 2017). For both, the best-fitting substitution model was inferred to be the general time reversible model with empirical base frequencies and two FreeRate categories for the spike (GTR+F+R2) and four FreeRate categories for the genome (GTR+F+R4) (Yang 1995, Soubrier et al. 2012). Trees were visualized using the ggtree package (Yu et al. 2017).

Sieve analysis methods

We performed two types of sieve analyses: global sieve analyses, comparing the distance of sequences from the reference sequence, Wuhan-Hu-1, between the unvaccinated and vaccinated groups, and local sieve analyses, which looked at amino acid sites in the spike individually and tested whether there were differences in amino acids found in sequences sampled from unvaccinated and vaccinated participants (Edlefsen et al. 2015).

Global sieve analyses

Pairwise distances between tips in the phylogenetic tree were extracted with the cophenetic function in the ape package in (Paradis et al. 2004) R. Hamming distances were calculated using the Wuhan-Hu-1 sequence as the reference, counting sites with gaps (–) relative to the reference due to the importance of deletions in some variants, but ignoring those with unknown amino acid (X).

Local sieve analyses

For each polymorphic amino acid site in spike, we compared the probability of mismatch to the Wuhan-Hu-1 residue in the vaccinated group with the probability in the unvaccinated group, using the Z_1^A test statistic and permutation procedure for unadjusted P values from Gilbert et al. (2008). Since some sites have little amino acid variability, we used Tarone's modified Bonferroni procedure to adjust for multiple testing, computing the minimum achievable significance level, α_i^*, based on Fisher's exact test (Gilbert 2005, Rolland et al. 2011). This effectively prescreens out sites that are highly conserved, which can provide extra power to identify sites where there is sufficient diversity for hypothesis testing.

Antibody escape scores

Antibody escape scores were defined as described in Magaret et al. (2024). They are defined using complex structures of SARS-CoV-2 and antibodies available in the Protein Data Bank (PDB) ($n=274$ on 4 May 2022). For each PDB complex, epitope sites were defined as antigen sites that are in contact with the antibody in the antigen–antibody complex (i.e. all sites that have nonhydrogen atoms within 4 Å of the antibody). The interaction between an epitope site i and the antibody is defined as the weight w_i:

$$w_i=1/2\left(n_{c_i}/\langle n_c\rangle+n_{nb_i}/\langle n_{nb}\rangle\right), \quad (1)$$

in which n_{c_i} is the number of contacts with the antibody (i.e. the number of nonhydrogen antibody atoms within 4 Å of the site); n_{b_i} is the number of neighboring antibody residues; $\langle n_c\rangle$ is the mean number of contacts n_c, and $\langle n_{nb}\rangle$ is the mean number of neighboring antibody residues n_{nb} across all epitope sites. A weight of 1.0 is attributed to the average interaction across all epitope sites. Neighboring residue pairs were identified by Delaunay tetrahedralization of side chain centers of residues (C_a is counted as a side chain atom, and pairs further than 8.5 Å were excluded).

The epitope distance between a virus sequence X and a reference sequence R (corresponding to the vaccine insert) was defined as the weighted mean of the distance between all epitope sites:

$$D(R,X)=\sum_i w_i\cdot \mathrm{Dist}(X_i,R_i)/\sum_i w_i, \quad (2)$$

$$\mathrm{Dist}(X_i,R_i)=1/2\cdot[\mathrm{Sim}(R_i,R_i)+\mathrm{Sim}(X_i,X_i)]-\mathrm{Sim}(X_i,R_i), \quad (3)$$

in which $\mathrm{Dist}(X_i,R_i)$ is the sequence distance between epitope site i; $\mathrm{Sim}(X_i,R_i)$ is the amino acid similarity according to the BLOSUM62 matrix. The distance between amino acid pairs includes insertion/deletion and glycosylation (match, 0; mismatch, –13, the worst substitution).

Epitope distances calculated for the 274 antibodies are compiled to define summary measures for 14 representative clusters of antibody footprints.

Results

Sample population characteristics and temporal/spatial distributions of emerging SARS-CoV-2 variants

Of 7911 individuals enrolled in the EPIC cohort, 5246 had tested positive for SARS-CoV-2 and 1327 had samples that had undergone whole-genome sequencing. Among the 595 who met the inclusion criteria (first testing positive for SARS-CoV-2 after general vaccine availability in December 2020, being unvaccinated or fully vaccinated, and having a good-quality whole-genome sequence available), 58% were unvaccinated at the time of infection (Table 1). Individuals in the unvaccinated group were on average younger than vaccinated individuals, possibly due to the staggered rollout of vaccine availability and/or participant risk perception. Unvaccinated individuals were more likely to be hospitalized with acute COVID-19 (36.3% versus 11.7%, $P < .0001$). Unvaccinated and vaccinated participants had similar Charlson comorbidity index scores. Samples were generally taken within a week of first positive test (median: 3 days; interquartile range, IQR: 0–8 days), with sequences tending to be sampled slightly later in the vaccinated group (median: 5, IQR: 1–8) versus the unvaccinated group (median: 2, IQR: 0–8).

Sequences were typed as one WHO VOC/variant of interest (VOC/VOI) or whether they contained the D614G mutation in the spike. WHO-named variant lineages accounted for 66.2% ($n = 394$) of the infections considered in this study, with all VOCs identified in at least two individuals (Fig. 1, Supplementary Fig. S2). B.1.2 was the most frequently identified nonvariant lineage ($n = 88$, 14.8%).

Table 1. Characteristics of EPICC participants included in analyses.

	Unvaccinated ($N = 347$)	Vaccine breakthrough ($N = 248$)	Total ($N = 595$)	P value[a]
Gender, *n* (%)				.21
Female	151 (43.5)	95 (38.3)	246 (41.3)	
Male	196 (56.5)	153 (61.7)	349 (58.7)	
Race and ethnicity, *n* (%)				.01
Black	39 (11.2)	29 (11.7)	68 (11.4)	
Hispanic or Latino	93 (26.8)	39 (15.7)	132 (22.2)	
Others	30 (8.6)	22 (8.9)	52 (8.7)	
White	185 (53.3)	158 (63.7)	343 (57.6)	
Age group, years, *n* (%)				<.01
<18	62 (17.9)	8 (3.2)	70 (11.8)	
18–44	144 (41.5)	138 (55.6)	282 (47.4)	
45–64	107 (30.8)	74 (29.8)	181 (30.4)	
≥65	34 (9.8)	28 (11.3)	62 (10.4)	
Charlson Comorbidity Index, *n* (%)				.58
0	199 (57.3)	155 (62.5)	354 (59.5)	
1–2	83 (23.9)	49 (19.8)	132 (22.2)	
3–4	38 (11.0)	24 (9.7)	62 (10.4)	
>5	27 (7.8)	20 (8.1)	47 (7.9)	
Severity, *n* (%)				<.01
Hospitalized	126 (36.3)	29 (11.7)	155 (26.)	
Outpatient	221 (63.7)	219 (88.3)	440 (73.9)	
Variant, *n* (%)				<.01
Alpha	44 (12.7)	12 (4.8)	56 (9.4)	
Beta	1 (0.3)	1 (0.4)	2 (0.3)	
Delta	82 (23.6)	136 (54.8)	218 (36.6)	
Epsilon	11 (3.2)	3 (1.2)	14 (2.4)	
Eta	1 (0.3)	0 (0.0)	1 (0.2)	
Gamma	2 (0.6)	4 (1.6)	6 (1.0)	
Iota	5 (1.4)	0 (0.0)	5 (0.8)	
Mu	0 (0.0)	1 (0.4)	1 (0.2)	
Omicron	7 (2.0)	83 (33.5)	90 (15.1)	
Other	193 (55.6)	8 (3.2)	201 (33.8)	
Zeta	1 (0.3)	0 (0.0)	1 (0.2)	
Site, *n* (%)				<.01
BAMC	122 (35.2)	41 (16.5)	163 (27.4)	
CRDAMC	14 (4.0)	0 (0.0)	14 (2.4)	
ATAMMC	15 (4.3)	17 (6.9)	32 (5.4)	
MAMC	38 (11.0)	47 (19.0)	85 (14.3)	
NMCP	12 (3.5)	3 (1.2)	15 (2.5)	
NMCSD	13 (3.7)	4 (1.6)	17 (2.9)	
TAMC	11 (3.2)	18 (7.3)	29 (4.9)	
WAMC	2 (0.6)	8 (3.2)	10 (1.7)	
WBAMC	17 (4.9)	11 (4.4)	28 (4.7)	
WRNMC	103 (29.7)	99 (39.9)	184 (33.9)	

[a] $n \times k$ Fisher's exact test.

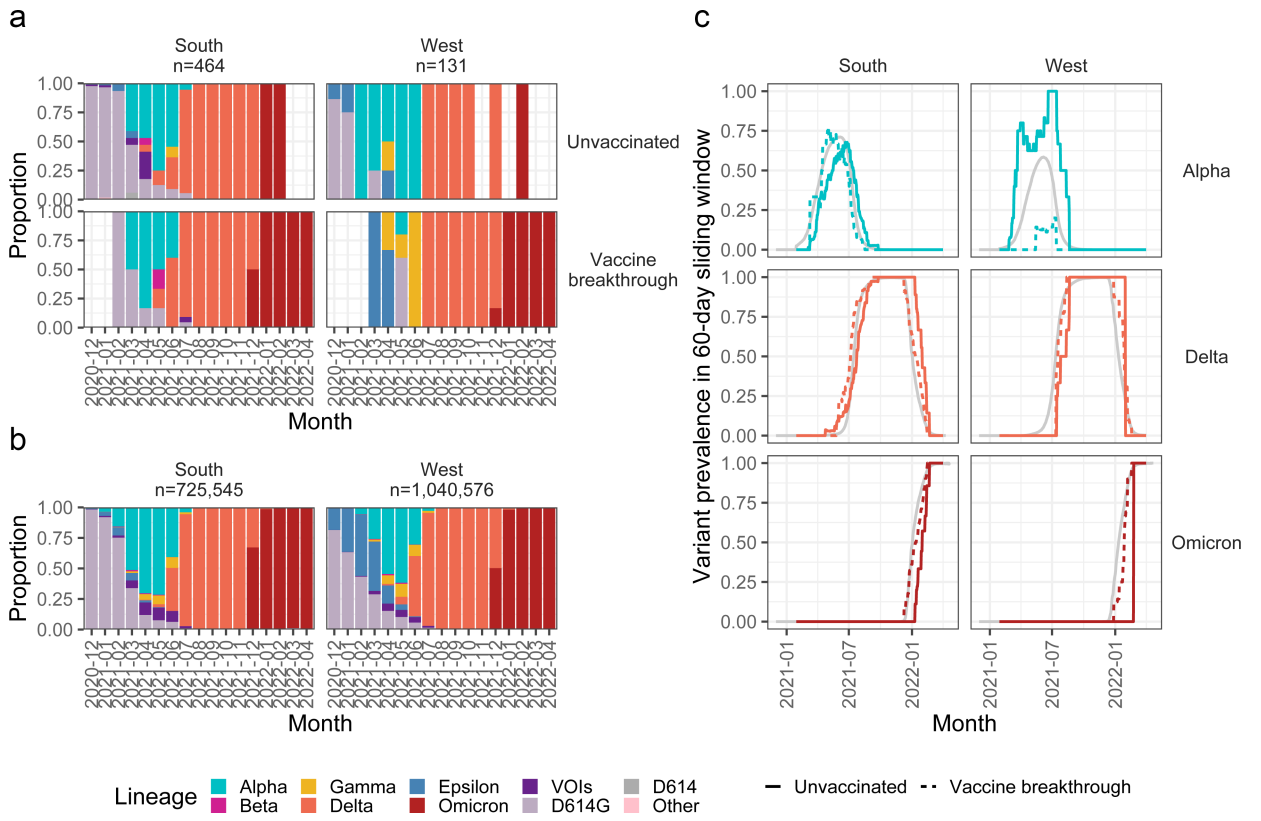

Figure 1. Circulation of variants in the South and West census bureau regions of the USA during the EPICC study. (a) The proportion of sequences broken down by variant from the EPICC study. WHO-named VOIs are included in a single category, with nonvariant spike D614G-containing lineages also separated out. (b) The proportion of sequences by variant in GISAID for the same period as the EPICC study, downloaded on 28 April 2022. (c) Variant prevalence in a 60-day sliding window for unvaccinated (solid line) and vaccine breakthrough (dashed line) groups for VOCs with at least five sequences per group. Gray lines denote the rolling averages from GISAID over the same period.

Since there were differing circulation patterns of variants, particularly for Epsilon and Iota, across the USA during the study period, for analysis purposes, we split samples according to the South ($n = 464$) and West ($n = 131$) census regions (Fig. 1 and Supplementary Fig. S3). Trends were less clear in the West region, with sparser collection and lower overall numbers that did not reflect the patterns seen in matched data from GISAID (Fig. 1a and Supplementary Fig. S3). Thus, we focused on data from the South region, where the sample appeared to be representative of overall circulation patterns, and we had more power to identify viral genomic differences between unvaccinated and vaccinated individuals.

The prevalence of variants in the South correlated with the trends seen in the context sequences downloaded from GISAID (Fig. 1a and b), with Alpha first emerging on the background of D614G-containing lineages, then being replaced by Delta, and then by Omicron. In the monthly data, we saw an earlier peak of Alpha in the vaccine breakthrough group (Fig. 1a), with a suggestion of a slight shift left in the 60-day sliding window prevalence of Alpha, Delta, and Omicron (Fig. 1c). However, this shift was not significant for any of the variants when tested by a logistic regression of emergence with the vaccination group included as a covariate (Supplementary Fig. S4). Also interesting was that the 60-day sliding window prevalence from the GISAID context data more resembled the vaccine breakthrough group than the unvaccinated group (Fig. 1c).

SARS-CoV-2 sequences from postvaccine infections intermingle with those from unvaccinated infections in phylogenies

One of the ways a global sieve effect by strain or variant would manifest is by sequences from unvaccinated and vaccine breakthrough infections clustering separately in the phylogenetic tree. Although such an extreme separation would be unlikely, we reconstructed the phylogenetic tree using IQ-TREE (Nguyen et al. 2014), with nucleotide sequences spanning the genome (Supplementary Fig. S5) and the spike (Fig. 2). As expected, sequences were broadly clustered by WHO variant for both the full genome and spike trees although the spike tree had reduced resolution. Variants that emerged earlier in the pandemic, such as Alpha and Gamma, were closer to the root, set at Wuhan-Hu-1, than Delta and Omicron that emerged later.

Within variants, unvaccinated and vaccine breakthrough sequences were intermingled with no clear subclade structure. To see if there were differences in standing genetic diversity within variant clades, we considered the pairwise distance between matched variant tips in the tree for both groups (Fig. 2 and Supplementary Fig. S5). We saw differences in the distributions using the Wilcoxon rank sum test, with the vaccine breakthrough group showing higher pairwise diversity for Alpha and D614G-containing sequences and the unvaccinated group having a slightly higher median pairwise difference Omicron sequences. The greatest difference was seen in the D614G group; this is, by definition, a

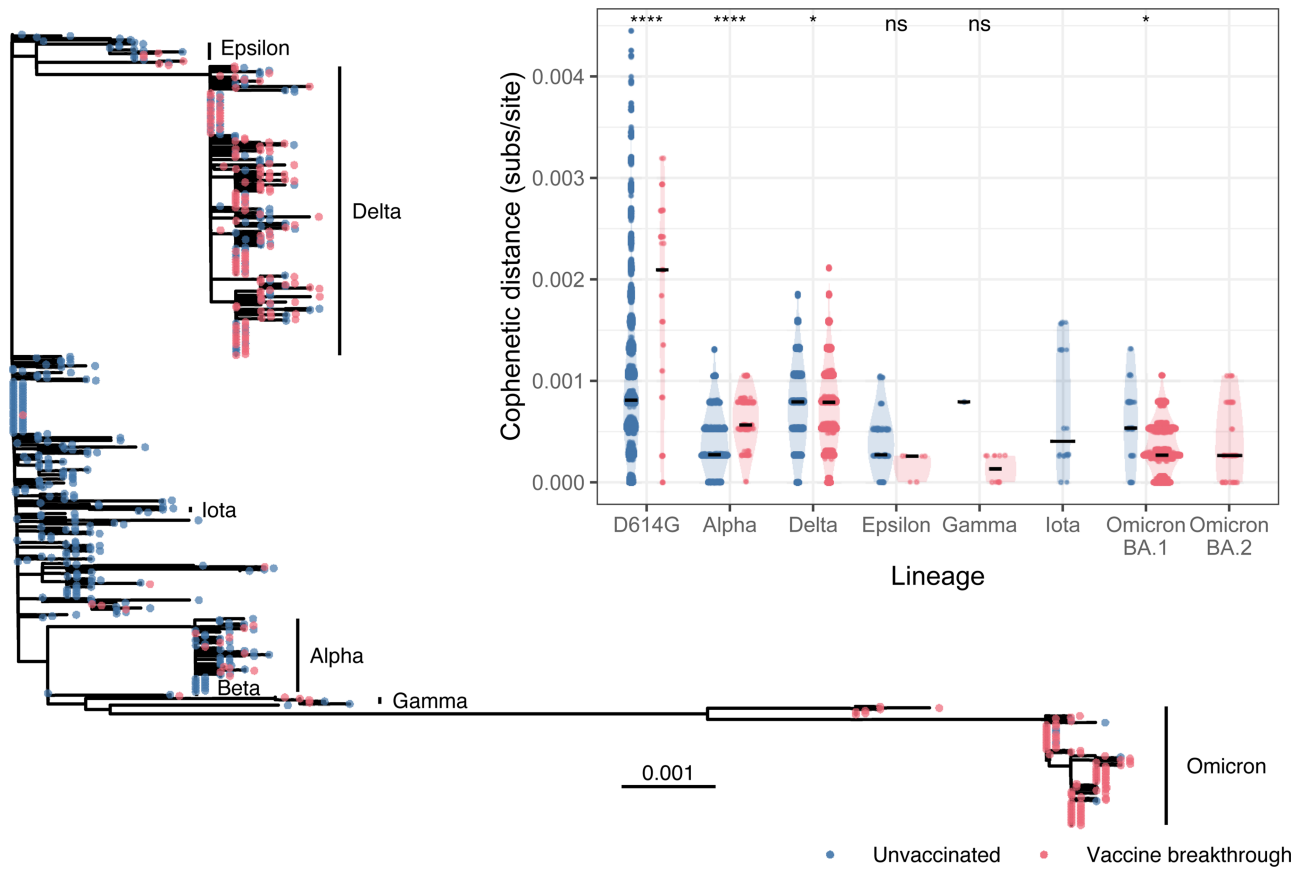

Figure 2. Sequences from unvaccinated and vaccine breakthrough infections are intermingled in the phylogenetic tree. Maximum likelihood tree reconstructed from spike nucleotide sequences with IQ-TREE and rooted by Wuhan-Hu-1; tip points are slightly jittered for readability. Inset: comparison of pairwise distances between tips of the same lineage for the unvaccinated and vaccine breakthrough cases. Asterisks give significance levels for *P* values: **** $P < .0001$; *** $P < .001$; ** $P < .1$, * $P < .05$; ns, not significant.

heterogeneous group of lineages with different ancestral histories, which also shows a heavy skew toward sequences from early infections in unvaccinated participants (Supplementary Fig. S3). Thus, the difference also likely captured temporal skew as the virus accumulated mutations and had varying effective population size through time.

Evaluation of divergence from Wuhan-Hu-1 in the spike region indicates no significant differences between vaccinated and unvaccinated infections

To investigate whether there were differences between the two groups at the amino acid site level, we considered the divergence of the spike protein away from the Wuhan-Hu-1 reference sequence using the Hamming distance (Fig. 3). We hypothesize that sequences from vaccine breakthrough infections will be further from the Wuhan-Hu-1-like vaccine insert. Since deletions have been identified in several variants, we counted deletions on a site-wise basis but ignored any ambiguous positions. In the unvaccinated group, we saw a bimodal distribution with the D614G-containing lineages in the first peak and a second peak containing the more distant VOCs (Fig. 3a). There was a similar peak of variants in the vaccine breakthrough group, with a second peak of Omicron sequences. Due to the succession of variants, we also split the data by quarter (Fig. 3b). We found no significant differences between the two groups within time periods, but there was significant divergence away from the reference sequence between subsequent quarters for all except Quarter 2 to Quarter 3 in 2021, during which Delta became fixed within the USA (Fig. 1b).

We also looked at the shift in divergence from the reference in a continuous manner using a linear regression (Fig. 3c). The best-fitting model included a unique intercept and slope for each vaccination group ($P < .0001$ for both). However, this was also influenced by the sampling biases over time, whereby there were more unvaccinated individuals sampled during winter 2020–21 and more vaccine breakthrough infections toward the end of the study (Supplementary Fig. S2). When we considered only Delta sequences—i.e. time and variant matched—we saw no such large effects, with the linear model fit by stepwise regression only containing the vaccination group, with sequences from vaccinated individuals having an average Hamming distance of 0.36 less than those from unvaccinated individuals ($P = .0064$).

Evaluation for site-level sieve effects does not show robust vaccine-specific diversification

Finally, we considered local sieve effects at the amino acid level. For each site in the spike, we calculated the number of amino acid mismatches with Wuhan-Hu-1 to identify sites that could distinguish between the unvaccinated and vaccine breakthrough groups. Overall, we found 45 sites that were significantly more diverse in the vaccination breakthrough group than in the unvaccinated group, of which 17 were in the N-terminal domain (NTD) and 17 were in the receptor binding domain (RBD) (Fig. 4a). All 45

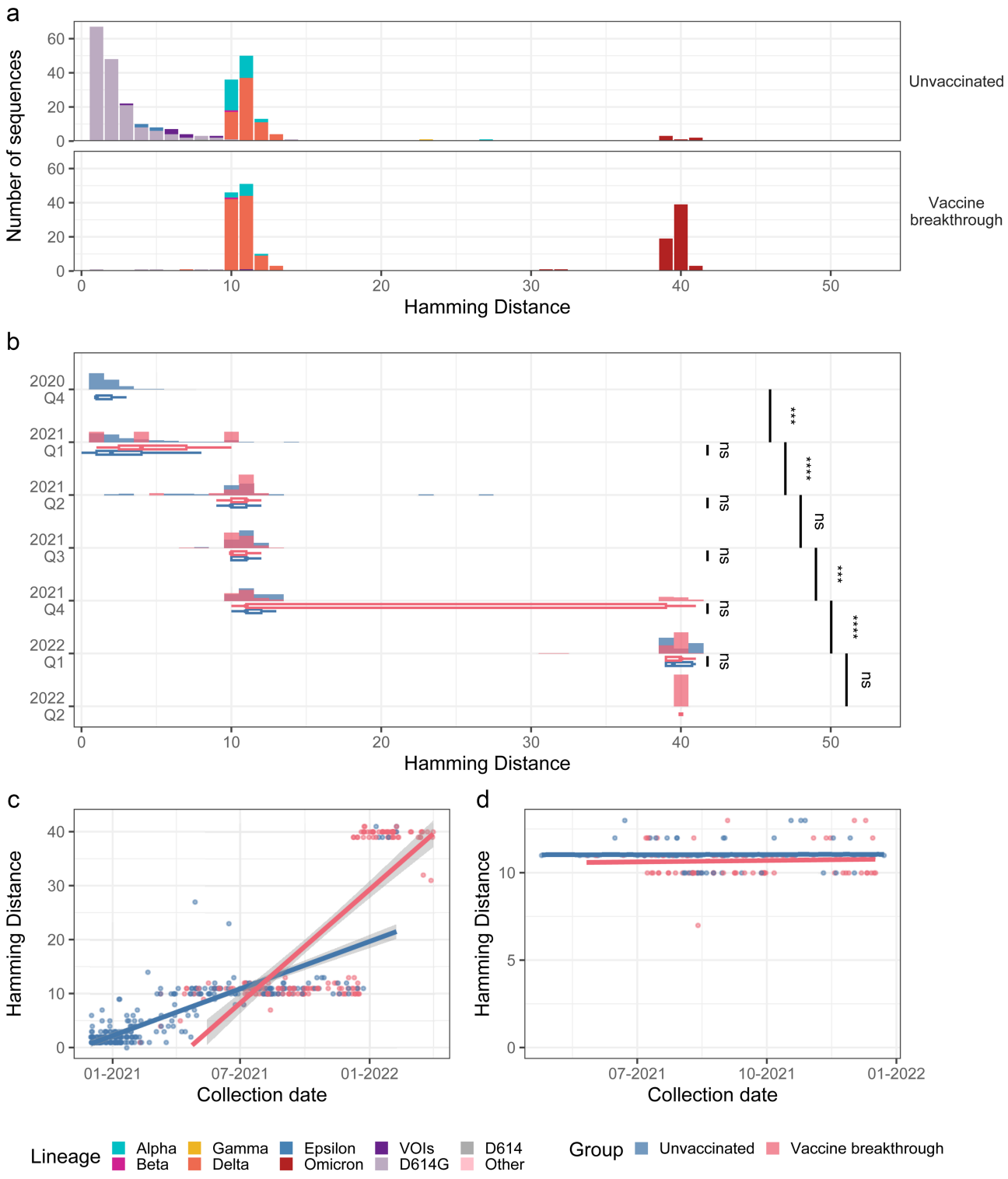

Figure 3. Spike divergence for sequences sampled in the South census region, calculated using the Hamming distance from Wuhan-Hu-1. (a) Overall distribution of Hamming distance split by vaccination group, broken down by lineage. (b) Comparison of Hamming distance distributions over time. Differences between vaccination groups and between time points compared with the Wilcoxon rank sum test and Bonferroni adjusted for multiple testing. (c) Linear regression of the Hamming distance and sample collection date. (d) As in (c), but for Delta sequences only. Asterisks give significance levels for *P* values: **** $P < .0001$; *** $P < .001$; ** $P < .01$; * $P < .05$; ns, not significant.

mutations are characteristic of at least one named WHO VOC/VOI, and 18 are characteristic of at least two different variants (Table 2).

However, these effects disappeared when time windows were analyzed (Fig. 4b and Supplementary Fig. S6). Considering the data broken down by quarter in Fig. 4b, we saw no signature sites differentiating the two groups until Quarter 4 in 2021. At this point, a constellation of mutations appeared in consistent frequency in the vaccination group. These mutations were all associated with

Table 2. List of signature sites in the spike and the WHO-named variants in which they are known to be characterizing.

Site	Mutation	Domain	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Mu	Omicron	Theta	Zeta
19	T19R	NTD			✓										
	T19I	NTD											✓		
67	A67V	NTD					✓						✓		
69	H69-	NTD	✓				✓						✓		
70	V70-	NTD	✓				✓						✓		
95	T95I	NTD			✓				✓	✓		✓	✓		
142	G142D	NTD			✓					✓			✓		
	G142-	NTD												✓	
143	V143-	NTD											✓	✓	
144	Y144-	NTD	✓				✓						✓		
	Y144S	NTD										✓			
145	Y145N	NTD										✓			
	Y145-	NTD											✓		
156	E156G	NTD			✓										
157	F157-	NTD			✓										
158	R158-	NTD			✓										
211	N211-	NTD											✓		
212	L212I	NTD											✓		
ins214	ins214EPE	NTD											✓		
339	G339D	RBD											✓		
346	R346K	RBD										✓			
371	S371L	RBD											✓		
	S371F	RBD											✓		
373	S373P	RBD											✓		
375	S375F	RBD											✓		
417	K417N	RBD		✓									✓		
	K417T	RBD						✓							
440	N440K	RBD											✓		
446	G446S	RBD											✓		
452	L452R	RBD			✓	✓				✓			✓		
	L452Q	RBD									✓				
477	S477N	RBD							✓				✓		
478	T478K	RBD			✓								✓		

(continued)

Table 2. (Continued)

Site	Mutation	Domain	Alpha	Beta	Delta	Epsilon	Eta	Gamma	Iota	Kappa	Lambda	Mu	Omicron	Theta	Zeta
484	E484K	RBD		✓			✓	✓	✓			✓		✓	✓
	E484Q	RBD								✓					
	E484A	RBD											✓		
493	Q493R	RBD											✓		
496	G496S	RBD											✓		
498	Q498R	RBD											✓		
501	N501Y	RBD	✓	✓				✓				✓	✓	✓	
505	Y505H	RBD											✓		
547	T547K												✓		
655	H655Y							✓					✓		
679	N679K												✓		
681	P681H		✓									✓	✓	✓	
	P681R				✓					✓					
764	N764K												✓		
796	D796Y												✓		
856	N856K												✓		
950	D950N				✓							✓			
954	Q954H												✓		
969	N969K												✓		
981	L981F												✓		

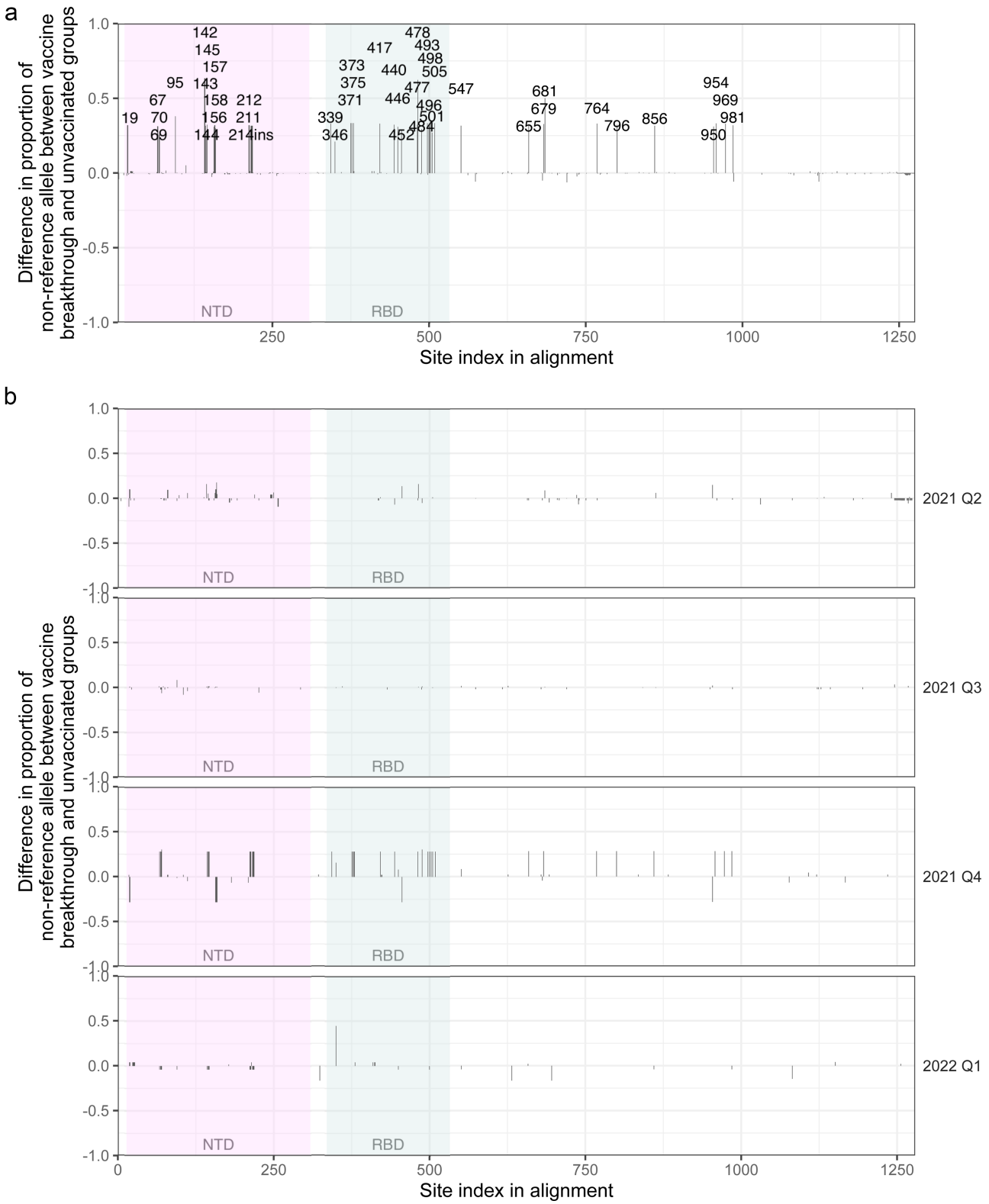

Figure 4. Local sieve analysis of site diversity for (a) overall in the South census region and (b) broken down by year quarter. A proportion >0 shows higher diversity from Wuhan-Hu-1 in the vaccine breakthrough group, whereas proportion <0 shows higher diversity in the unvaccinated group. Significant sites from the permutation test after adjustment for multiple testing are labeled by their Wuhan-Hu-1 index (214ins = insertion after site 214). Only quarters with more than five sequences available in both groups were analyzed.

Omicron and were significant before adjusting for multiple testing. There was a Delta sequence in the vaccine breakthrough group that also had nonsynonymous mutations at two sites (70 and 484) with Omicron-specific mutations. However, both substitutions were different amino acids compared to Omicron–V70I (Omicron: del70) and E484Q (Omicron: E484A).

We also saw slightly different patterns looking by month (Supplementary Fig. 6). Although not significant after adjusting for multiple testing, mutations corresponding to the characterizing sites in Alpha were more prevalent in the vaccinated group in April 2021 and then switched to being in higher frequency in the unvaccinated group in May. Then, Delta characterizing mutations appeared in higher frequency in the vaccinated group in June, but once Delta dominated, there was very little diversity between the two groups, which was reflected in the tight Hamming distance distributions (Fig. 3a).

To better account for antibody escape, we calculated escape scores or epitope distances for 14 representative clusters of antibody footprints on SARS-CoV-2 spike (2 in the NTD and 12 in the RBD), thereby summarizing multiple mutations that would simultaneously affect the interaction with an antibody. We found evidence of antibody escape with larger epitope distances in the vaccinated group during the fourth quarter of 2021, driven by differences found in December 2021 (Supplementary Fig. S7). Significant differences were found for 12 of the 14 epitope clusters; the greatest differences in mean PDB escape score (vaccine–placebo) were seen for PDB14 in the NTD (overall difference: 1.25, December 2021: 2.04; $P < .0001$) and PDB3 in the RBD (overall difference: 0.71, December 2021: 1.19; $P < .0001$). One of the highest weighted sites in PDB14 is 211, which is deleted in Omicron, whereas PDB3 is defined by several epitope sites (446, 493, 496, 498, and 501) that are mutated in Omicron.

Discussion

We analyzed SARS-CoV-2 sequences collected in the EPICC cohort between December 2020 and April 2022 to evaluate whether there were differences in sequences depending on the vaccination status of the participant. We saw some evidence suggesting that substitutions were first visible in vaccine breakthrough infections. However, across analyses, differences between the vaccinated and unvaccinated groups were typically not statistically significant.

Convergent evolution of SARS-CoV-2 with substitutions repeatedly observed at the same sites in the RBD has been noted since the appearance of variants (Martin et al. 2021, Valério et al. 2022). It is obvious that the convergent evolution of SARS-CoV-2 reflects the strong pressure exerted by spike-specific and particularly RBD-specific antibody responses (Starr et al. 2021, Cao et al. 2023). Yet, in our study, it was not possible to associate the evolution of variants directly with a vaccine-mediated effect. This can be due to multiple reasons, which serve as key methodological insights for future observational studies examining sieve effects.

First, we recognize that sieve analysis should be ideally conducted in the context of a VE trial where vaccine and placebo recipients are randomized and blinded (Gilbert et al. 2008, Rolland et al. 2011, Edlefsen et al. 2015). In observational studies, it is necessary to engineer a control (unvaccinated) group through stratification of time and space, which limits the statistical power to find small sieve effects. Although we had data from December 2020 to April 2022, not all months had at least five sequences in each vaccination status group. In addition, that the vaccine was already known to have good efficacy during the time of our study would have impacted behavior and decisions about risk of infection postvaccination in the absence of a placebo over the course of our study. Hence, our analysis of sequences from vaccine recipients using as a comparison an engineered control group corresponding to the unvaccinated participants in the cohort is an imperfect solution, although the two groups were broadly comparable.

Second, there was an imbalance in the distribution of the two groups that reflected the uptake of vaccination. The first sequences were obtained in December 2020, and sequences collected over the first 6 months were predominantly isolated from unvaccinated participants due to the limited availability of vaccines at that time; according to numbers from the CDC, around 50% of the population was fully vaccinated by July 2021 (CDC 2023). In contrast, sequences collected in the last 6 months of our study (until April 2022) corresponded mostly to vaccinated participants, as few individuals remained unvaccinated at that time. The rollout of the vaccine in the USA was also tied to risk of exposure (e.g. essential workers) and risk of infection outcome (e.g. immunocompromised status and age). Thus, we had a time-dependent distribution of participants in the two groups in addition to the time-dependent distribution of variants. Both factors limit our ability to detect genetic variation between the vaccinated and unvaccinated groups.

Third, our study population is drawn from a hybrid immunological landscape corresponding to different combinations of natural infection (possibly multiple times with different variants), partially and fully immunized individuals. Moreover, the massive scale of the pandemic with large numbers of infections and numerous infections remaining undiagnosed mean that it is likely that exposure rates overwhelmed sieve effects on viral acquisition. Since both SARS-CoV-2 infection and vaccination induce selective pressures that lead to similar substitutions focused on the spike, it is difficult to dissociate an effect that would be solely due to vaccination. While we found evidence of substitutions and larger escape scores appearing sooner among vaccinated individuals, it was not possible to consider diversification as a vaccine-specific effect in the context of our study (which had no randomization or blinding). Using antibody escape scores, which combine effects across each antibody footprint on SARS-CoV-2, there were larger epitope distances among vaccinated participants, specifically in December 2021. Importantly, December 2021 corresponded to the shift toward Omicron variants, suggesting that there was a vaccine effect promoting an earlier shift to Omicron among the vaccinated. (In a prior study on HIV-1 breakthrough infections, we showed that larger epitope distances associated with diminished prevention efficacy (Juraska et al. 2024)).

Overall, we observed differences linked to variants, with shifts happening quickly—faster than it can be detected by our temporal resolution when summarizing by quarter or month, making it difficult to ascribe differences to the sole impact of vaccination. During our study period, SARS-CoV-2 changes appeared in large evolutionary jumps, with the rapid accrual of a constellation of mutations rather than a gradual accumulation. The mechanisms behind these saltation events and where new variants arose are still uncertain (Corey et al. 2021, Harari et al. 2022). Here, we considered that the phylogenetic distance between successive variants cannot be attributed to a vaccine-only effect; hence, when accounting for the phylogenetic lineage, differences between the vaccinated and unvaccinated groups were not significant. Our results may partly be a consequence of the limited range of diversity among variants seen in the USA. As such, the sieve analysis conducted in the ENSEMBLE VE trial, which evaluated breakthrough infections in Latin America, the USA, and South Africa, only found evidence of a sieve effect in Latin America (Magaret et al. 2024). The sieve signal in Latin America was mostly linked to spike mutations defining the Lambda variant; interestingly, mutations at some of these sites (252, 484, 490, and 501) have appeared subsequently in different SARS-CoV-2 lineages.

This finding reflects that SARS-CoV-2 variants over time illustrate remarkable patterns of convergent evolution in the RBD. The limited set of RBD sites that have mutated across variants since the beginning of the pandemic is due to an antibody response heavily focused toward the RBD (Starr et al. 2021, Cao et al. 2023) (this also manifested in our study with differences between groups better defined when considering antibody epitope footprints than individual sites on spike). Hence, our results reflect a context of mass vaccination with a globally circulating virus with a high attack rate where vaccine-driven effects and infection-driven effects cause the same mutations and are difficult to untangle.

Acknowledgements

We gratefully acknowledge the authors and originating and submitting laboratories of the sequences from GISAID's EpiCov Database on which this research is based. We thank the members of the EPICC COVID-19 Cohort Study Group for their many contributions in conducting the study and ensuring effective protocol operations. The following members were all closely involved with the design, implementation, and oversight of the study:

Brooke Army Medical Center, Fort Sam Houston, TX: J. Cowden; M. Darling; S. DeLeon; D. Lindholm; A. Markelz; K. Mende; S. Merritt; T. Merritt; N. Turner; and T. Wellington.

Carl R. Darnall Army Medical Center, Fort Hood, TX: S. Bazan and P.K. Love.

Fort Belvoir Community Hospital, Fort Belvoir, VA: N. Dimascio-Johnson; N. Elnahas; E. Ewers; K. Gallagher; C. Glinn; U. Jarral; D. Jennings; D. Larson; K. Reterstoff; A. Rutt; A. Silva; and C. West.

Henry M. Jackson Foundation, Inc., Bethesda, MD: P. Blair; J. Chenoweth; and D. Clark.

Madigan Army Medical Center, Joint Base Lewis McChord, WA: J. Bowman; S. Chambers; C. Colombo; R. Colombo; C. Conlon; K. Everson; P. Faestel; T. Ferguson; L. Gordon; S. Grogan; S. Lis; M. Martin; C. Mount; D. Musfeldt; D. Odineal; M. Perreault; W. Robb-McGrath; R. Sainato; C. Schofield; C. Skinner; M. Stein; M. Switzer; M. Timlin; and S. Wood.

Naval Medical Center Portsmouth, Portsmouth, VA: S. Banks; R. Carpenter; L. Kim; K. Kronmann; T. Lalani; T. Lee; A. Smith; R. Smith; R. Tant; and T. Warkentien.

Naval Medical Center San Diego, San Diego, CA: C. Berjohn; S. Cammarata; N. Kirkland; D. Libraty; R. Maves; and G. Utz.

Tripler Army Medical Center, Honolulu, HI: C. Bradley; S. Chi; R. Flanagan; A. Fuentes; M. Jones; N. Leslie; C. Lucas; C. Madar; K. Miyasato; and C. Uyehara.

Uniformed Services University of the Health Sciences, Bethesda, MD: H. Adams; B. Agan; L. Andronescu; A. Austin; B. Barton, D. Becher, C. Broder; T. Burgess; C. Byrne; K Chung; J. Davies; C. English; N. Epsi; C. Fox; M. Fritschlanski; A. Hadley; P. Hickey; E. Laing; C. Lanteri; J. Livezey; A. Malloy; A. Michel, R. Mohammed; C. Morales; P. Nwachukwu; C. Olsen; E. Parmelee; S. Pollett; S. Richard; J. Rothenberg, J. Rozman; J. Rusiecki; D. Saunders; E. Samuels; M. Sanchez; A. Scher; M. Simons; A. Snow; K. Telu; D. Tribble; M. Tso; L. Ulomi; and M. Wayman, N. Hockenbury.

US Air Force School of Aerospace Medicine, Dayton, OH: T. Chao; R. Chapleau; M. Christian; A. Fries; C. Harrington; V. Hogan; S. Huntsberger; K. Lanter; E. Macias; J. Meyer; S. Purves; K. Reynolds; J. Rodriguez; and C. Starr.

US Coast Guard, Washington, DC: J. Iskander and I. Kamara.

Womack Army Medical Center, Fort Bragg, NC: B. Barton; D. Hostler; J. Hostler; K. Lago; C. Maldonado; and J. Mehrer.

William Beaumont Army Medical Center, El Paso, TX: T. Hunter; J. Mejia; R. Mody; J. Montes; R. Resendez; and P. Sandoval.

Walter Reed National Military Medical Center, Bethesda, MD: I. Barahona; A. Baya; A. Ganesan; N. Huprikar; and B. Johnson.

Walter Reed Army Institute of Research, Silver Spring, MD: S. Peel.

Supplementary data

Supplementary data is available at *VEVOLU* Journal online.

Conflict of interest: The views expressed are those of the authors and do not reflect the official policy of the Uniformed Services University of the Health Sciences (USUHS), Department of the Army, Department of the Navy, the Department of the Air Force, the Department of Defense or the US Government, and the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. (HJF). The investigators have adhered to the policies for protection of human subjects as prescribed in 45 CFR 46. Drs Berjohn, Fries, Smith, Mody, Huprikar, Lindholm, Jones, Larson, Ewers; Ms Bazan; and Drs Saunders, Maldonado, Simons, Tribble, and Burgess are service members or employees of the US Government. This work was prepared as part of their official duties. Title 17 USC §105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 USC §101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.

Funding

This work was supported by awards from the Defense Health Program (HU00012020067) and the National Institute of Allergy and Infectious Disease (HU00011920111). The protocol was executed by the Infectious Disease Clinical Research Program (IDCRP), a Department of Defense (DoD) program executed by the USUHS through a cooperative agreement by the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. (HJF). This project has been funded in part by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health, under an interagency agreement (Y1-AI-5072). Work by the US Military HIV Research Program was supported by a cooperative agreement (WW81XWH-18-2-0040) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the US DoD.

Data availability

Sequences are available on GenBank under accession numbers OR611156–OR611708 and PP378952–PP379010.

References

Cao L, Lou J, Chan SY *et al.* Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance. *Nat Med* 2022;**28**:1715–22.

Cao Y, Jian F, Wang J *et al.* Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* 2023;**614**:521–29.

CDC. *COVID-19 Vaccination Age and Sex Trends in the United States, National and Jurisdictional*. 2023. https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Age-and-Sex-Trends-in-the-Uni/5i5k-6cmh/about_data (19 February 2024, date last accessed).

Corey L, Beyrer C, Cohen MS *et al*. SARS-CoV-2 variants in immunosuppressed individuals. *N Engl J Med* 2021;**385**:562–66.

Edlefsen PT, Rolland M, Hertz T *et al*. Comprehensive sieve analysis of breakthrough HIV-1 sequences in the RV144 vaccine efficacy trial. *PLoS Comput Biol* 2015;**11**:1–37.

Epsi NJ, Powers JH, Lindholm DA *et al*. A machine learning approach identifies distinct early-symptom cluster phenotypes which correlate with hospitalization, failure to return to activities, and prolonged COVID-19 symptoms. *PLoS One* 2023a; **18**:e0281272.

Epsi NJ, Richard SA, Lindholm DA *et al*. Understanding "Hybrid Immunity": comparison and predictors of humoral immune responses to severe acute respiratory syndrome coronavirus 2 infection (SARS-CoV-2) and coronavirus disease 2019 (COVID-19) vaccines. *Clin Infect Dis* 2023b;**76**:e439–49.

Feikin DR, Higdon MM, Abu-Raddad LJ *et al*. Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *Lancet* 2022;**399**:924–44.

Freed NE, Vlková M, Faisal MB *et al*. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol Methods Protoc* 2020;**5**:bpaa014.

Gilbert PB. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *J R Stat Soc Ser C Appl Stat* 2005;**54**:143–58.

Gilbert PB, Wu C, Jobes DV. Genome scanning tests for comparing amino acid sequences between groups. *Biometrics* 2008;**64**:198–207.

Haas EJ, Angulo FJ, McLaughlin JM *et al*. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *Lancet* 2021;**397**:1819–29.

Hacisuleyman E, Hale C, Saito Y *et al*. Vaccine breakthrough infections with SARS-CoV-2 variants. *N Engl J Med* 2021;**384**:2212–18.

Harari S, Tahor M, Rutsinsky N *et al*. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat Med* 2022;**28**:1501–08.

Juraska M, Bai H, DeCamp AC *et al*. Prevention efficacy of the broadly neutralizing antibody VRC01 depends on HIV-1 envelope sequence features. *Proc Natl Acad Sci USA* 2024; **121**:e2308942121.

Kalyaanamoorthy S, Minh BQ, Wong TKF *et al*. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;**14**:587–89.

Katoh K, and Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 2016;**32**:1933–42.

Khare S, Gurry C, Freitas L *et al*. GISAID's role in pandemic response. *China CDC Wkly* 2021;**3**:1049–51.

Lusvarghi S, Pollett SD, Nath Neerukonda S *et al*. SARS-CoV-2 BA.1 variant is neutralized by vaccine booster-elicited serum but evades most convalescent serum and therapeutic antibodies. *Sci Trans Med* 2022;**14**:eabn8543.

Magaret CA, Li L, deCamp AC *et al*. Quantifying how single dose Ad26.COV2.S vaccine efficacy depends on spike sequence features. *Nat Commun* 2024;**15**:1–22.

Marques AD, Sherrill-Mix S, Everett JK *et al*. SARS-CoV-2 variants associated with vaccine breakthrough in the Delaware valley through summer 2021. *mBio* 2021;**13**:e0378821.

Martin DP, Weaver S, Tegally H *et al*. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* 2021;**184**:5189–200.e7.

Mascola JR, Graham BS, Fauci AS. SARS-CoV-2 viral variants—tackling a moving target. *JAMA* 2021;**325**:1261–62.

Nguyen L-T, Schmidt HA, von Haeseler A *et al*. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2014;**32**: 268–74.

O'Toole Á, Scher E, Underwood A *et al*. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* 2021;**7**:1–9.

Paradis E, Claude J, Strimmer K *et al*. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004;**20**:289–90.

Richard SA, Epsi NJ, Lindholm DA *et al*. COVID-19 patient-reported symptoms using FLU-PRO Plus in a cohort study: associations with infecting genotype, vaccine history, and return to health. *Open Forum Infect Dis* 2022;**9**:ofac275.

Richard SA, Pollett SD, Fries AC *et al*. Persistent COVID-19 symptoms at 6 months after onset and the role of vaccination before or after SARS-CoV-2 infection. *JAMA Netw Open* 2023;**6**: e2251360.

Richard SA, Pollett SD, Lanteri CA *et al*. COVID-19 outcomes among US Military Health System beneficiaries include complications across multiple organ systems and substantial functional impairment. *Open Forum Infect Dis* 2021;**8**:ofab556.

Rolland M, and Gilbert PB. Sieve analysis to understand how SARS-CoV-2 diversity can impact vaccine protection. *PLoS Pathog* 2021;**17**:e1009406.

Rolland M, Tovanabutra S, DeCamp AC *et al*. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat Med* 2011;**17**:366–71.

Soubrier J, Steel M, Lee MSY *et al*. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol Biol Evol* 2012;**29**:3345–58.

Starr TN, Greaney AJ, Addetia A *et al*. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* 2021;**371**:850–54.

Tartof SY, Slezak JM, Fischer H *et al*. Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *Lancet* 2021;**398**:1407–16.

Thompson MG, Burgess JL, Naleway AL *et al*. Interim estimates of vaccine effectiveness of BNT162b2 and mRNA-1273 COVID-19 vaccines in preventing SARS-CoV-2 infection among health care personnel, first responders, and other essential and frontline workers—eight U.S. locations, December 2020–March 2011. *MMWR Morb Mortal Wkly Rep* 2021;**70**:495–500.

Valério M, Borges-Araújo L, Melo MN *et al*. SARS-CoV-2 variants impact RBD conformational dynamics and ACE2 accessibility. *Front Med Technol* 2022;**4**:1–13.

Wang W, Lusvarghi S, Subramanian R *et al*. Antigenic cartography of well-characterized human sera shows SARS-CoV-2 neutralization differences based on infection and vaccination history. *Cell Host Microbe* 2022;**30**:1745–1758.E7.

Yang Z. A space-time process model for the evolution of DNA sequences. *Genetics* 1995;**139**:993–1005.

Yu G, Smith DK, Zhu H *et al*. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;**8**:28–36.

Virus Evolution, 2024, 10(1), veae071 , DOI: https://doi.org/10.1093/ve/veae071, Advance Access Publication 2 September 2024, Research Article
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site–for further information please contact journals.permissions@oup.com.